Response to the public consultation on Guidelines 02/2026 on Anonymisation

Amir Kellou-Sidhoum, independent researcher · 13 October 2026

Comments on version 1.0, adopted on 7 July 2026. Paragraph and footnote numbers refer to that version. I write in a personal capacity. No one commissioned or funded this response.

Summary

The Guidelines already contain the right test for synthetic data. An inference counts when it relies on the individual's data and could not be obtained from data about the population at large (paragraph 71, footnotes 42 to 44), and paragraph 82 applies this to synthetic data and AI models. My comments concern how that test is run and read when a release holds one synthetic record per real person, such as a language-model "twin" of each survey respondent. My research on such releases points to two problems. A single re-identification rate, read on its own, can be wrong in both directions. And a test that finds nothing may have been unable to find anything. I ask the EDPB to:

1. **Require a matched baseline** (paragraph 71(b), footnote 43). Compare the given data with a version generated from the released attributes alone, attacked in exactly the same way, and never rely on a low success rate alone.
2. **Require a planted-trace test** (after paragraph 87). A test that finds nothing supports a criterion only if it was shown able to find something.
3. **Treat per-person synthetic records as linkable to the answers they were generated from** (paragraphs 65 and 85), and test distinct groups of questions separately.
4. **Treat random generation as a process, not a file** (footnote 43, paragraph 41). Compare several runs, and document enough to re-run the assessment.

Each proposal adds a few sentences, and a new example is drafted below. None of them changes the legal test in Article 4(1) and Recital 26 GDPR, or the assessment from each relevant entity's perspective that the Guidelines draw from C-413/23 P *EDPS v SRB*. They concern the evidence a controller relies on to show that the likelihood of identification is insignificant (paragraph 22).

About this response. I am an independent researcher working on privacy and the governance of AI systems. I audit published synthetic data for re-identification risk. I give principles and orders of magnitude here, not dataset-level figures. The findings come from a small number of releases and from test data I built with known ground truth. I offer them, in the spirit of paragraph 84, as reasons for the checks below, not as rates that hold elsewhere. Neither check is new. Comparing an attack's success with a reference point is established in the risk assessment of synthetic data, for example a control set of records not used to generate it (Giomi et al., PoPETs 2023) or the release of the raw data (Stadler, Oprisanu and Troncoso, USENIX Security 2022), and in attribute inference against trained models, where the reference is what can be imputed without the model (Jayaraman and Evans, ACM CCS 2022). Planting known records, often called canaries, and checking whether the process lets them through is established in the auditing of machine-learning models and generated text (Carlini et al., USENIX Security 2019; Jagielski, Ullman and Oprea, NeurIPS 2020; Meeus et al., ICML 2025). My proposals ask that the Guidelines make these checks expected where testing is relied on.

A proposed example: the twin of a survey respondent

The comments below rest on one case, which I propose adding after paragraph 82, following Example 19.

> **Example 19a:** A survey organisation, acting as anonymising controller, holds the answers of its panel respondents. For each respondent, it gives a language model the respondent's age, sex, region and education, together with some of the respondent's earlier answers, and asks the model to answer a new questionnaire as that person would. It publishes the generated answers, one record per respondent, with

> the four demographic attributes and without names. The organisation itself, users of the archive holding the original answers, and anyone who obtains a leaked copy of those answers could try to link each generated record to its respondent.
>
> To assess this, the organisation generates a second set of records for the same respondents from the four demographic attributes alone. It then attacks both sets with the same technique, the same original answers, the same set of candidate respondents and the same decision rule. If the first set is linked to its respondents clearly more often than the second, the generated records rely on each respondent's own data, beyond the released attributes (paragraph 71(b)). If both sets are linked well above chance and the difference between them is shown to lie within a margin fixed in advance, the records are still linkable and the No Linkage criterion is still violated, but the risk comes mainly from the released attributes, so removing or coarsening them is the first safeguard to test. If neither set is linked, the organisation should not rely on that result until it has checked that its test could have detected such a link (proposed paragraph 87a).

Comment 1. A matched baseline (paragraph 71(b) and (c), footnotes 43 and 44)

Text concerned. Under paragraph 71(b), reliance on the individual's data, or its absence, can be shown "by comparing the outcome of inference with two versions of the given data". Footnote 44, on paragraph 71(c), asks whether the given data yields "a significantly different 'posterior belief'".

Problem. The text does not require both versions to be attacked in the same way, and for per-person synthetic data the second version contains no record for the person.

Why it matters. Because paragraph 80 allows linkage to be seen as a form of inference, and paragraph 82 applies this to synthetic data, the comparison in paragraph 71(b) is also the natural way to read a linkage test as evidence of reliance. It does not replace the No Linkage criterion in paragraph 60. A success rate cannot be read on its own. It changes with the technique, the additional information, the set of candidate individuals and the decision rule, whatever the data. When the given data holds one synthetic record per person, A(D') in footnote 43 holds no record for that person, so a linkage test run on it has nothing to find. That comparison then shows reliance whenever a record exists, and cannot show what the generation step adds to what the released attributes already reveal. A second comparison, against records generated from the released attributes alone, isolates that contribution. It adds to the comparison with A(D') and does not replace it.

Evidence.

- On one release, the synthetic records led back to most respondents. A matched baseline did almost as well. Most of the linkage risk came from the released attributes, not from the further personal data given to the model.
- On another release, the rate was very low, and it did not separate from its baseline. That shows only that the test found no signal. Whether it could have found one is the question of Comment 2.
- On test data where I controlled whether any trace existed, a rate above chance reported a trace in the great majority of cases where there was none, because the released attributes alone already beat chance.

Proposed wording (add at the end of paragraph 71(b), or to footnote 43):

> Both versions should be tested with the same technique, the same additional information, the same set of candidate individuals and the same decision rule. Where the given data contains one synthetic record for each individual, the comparison should also use a version generated by the same process (same model and version, instructions and settings) for the same individuals from only the attributes released with the given data (a "matched baseline"). Where the relevant entity holds further attributes, a second baseline using them may be reported separately, but does not replace the first. Chance level is not an adequate baseline when the released attributes are informative. The difference between the two versions should be reported with its uncertainty, using a paired comparison over the same individuals, and examined across individuals as well as on average, since a small group of distinctive individuals may be

> linkable while the average rate is close to the baseline. A difference that is not statistically significant does not show that the two versions are equivalent; equivalence should be claimed only within a margin fixed before testing. A low success rate measured on one version alone should not be relied on to exclude a violation. A high success rate shows a risk whatever its source; the comparison shows where the risk comes from. The released attributes themselves remain subject to the No Record Isolation and No Linkage criteria.

In line with paragraph 87 and Example 20, results should report the share of correct matches at a low, stated rate of false matches, rather than accuracy alone.

Comment 2. A planted-trace test (paragraphs 87 and 95)

Text concerned. Paragraph 95: data that fulfil all three criteria "can be presumed anonymous". Paragraph 87 assesses a technique "on its ability to generate accurate results".

Problem. The Guidelines ask whether an attack succeeds, but not whether the controller's test could have detected a success in the first place.

Why it matters. Controllers will often conclude that a criterion is met because their test found nothing. Opinion 28/2024 (paragraph 55) already states that successful testing against widely known, state-of-the-art attacks "can only be evidence for the resistance to those attacks". A second limit applies even to known attacks. A test with low detection power finds nothing whether or not there is something to find.

Evidence. I built test data in which each synthetic record carried a trace of its person of known form and strength, and test data with no trace. The strongest check I tested kept false alarms low and found most records that copied part of their person's answers, and nearly all of them once a larger share was copied. It missed most records that reproduced the person's underlying tendencies without copying any answer: it found fewer than one in ten of the weakest such traces, and about half of the strongest. A stricter decision rule found none of them. These test data are my own, and real traces may be easier or harder to find. But the direction is clear. A clean audit can be weak evidence of absence, and without a planted-trace test the controller cannot know how weak.

Proposed wording (new paragraph 87a):

> A test that does not detect a violation supports a criterion only to the extent that it could have detected one. Where the assessment relies on testing, the anonymising controller should measure the detection power of the test, that is, the share of planted traces it detects at the chosen false-alarm rate. This can be done by introducing into the input data a small number of records or answers carrying a trace of known form and strength, running the full process, and checking whether the test finds them in the output (a "planted-trace test"). Planting the trace directly in the output checks only the detector, not the process. The strength of the trace should be expressed on a stated scale, such as the share of the individual's answers reproduced. Several forms should be tested, at least verbatim copying and reproduction of the individual's tendencies without copying. Detection power should be measured at the same decision threshold and false-alarm rate, on the same number of candidate individuals, as the actual test, and reported with a confidence interval. The documentation should state the weakest trace the test detects reliably, and which forms it could not detect. A test of unknown detection power should not be the sole basis for considering a criterion met.

And in paragraph 41: "This documentation of the anonymisation process (including for the testing of the supposedly anonymous datasets, and the measured detection power of that testing) makes it possible to demonstrate …".

Comment 3. Records derived from survey answers (paragraphs 65 and 85, Example 24)

Text concerned. Paragraph 65: the No Linkage criterion "may be met, for example, in the case of answers given to an opinion survey", absent traceable demographic data. Example 24 treats a 30-question customer survey as anonymous despite unique records, because "there is no way to link the answers" to the respondent.

Problem. This holds for survey answers on their own, but not once those answers are used to generate a synthetic record for each respondent, because that record is built to resemble the person's real answers.

Why it matters. Paragraph 65 itself sets the condition that the information is "not correlated to similar information about the same individual recorded in other contexts". A per-respondent synthetic record is such information by design. Anyone who holds or can obtain the real answers can try to link the two: the survey organisation, users of a data archive, or the recipient of a leaked file.

Evidence. In my audits, the link between a synthetic record and its respondent depended on which answer went with which question: shuffling that assignment removed it. Which questions were attacked also decided whether a rate could be read at all. On one release, one block of questions carried the linkage. On another block of the same questionnaire, no rate separated from its baseline. A single rate over the whole questionnaire hid that split.

Proposed wording (add at the end of paragraph 65):

> The opinion-survey example does not apply where a dataset derived from the same answers, such as synthetic records generated for each respondent, is available to a relevant entity who holds or can obtain the original answers. Such a derived dataset is generated to be correlated with similar information about the same individual, and should be tested accordingly.

Proposed wording (new bullet in paragraph 85):

> The content of the attributes. The criteria may be met for some groups of attributes and violated for others. Where the given data combines distinct groups of attributes, such as blocks of survey questions of different types, each group should be tested separately as well as together. The groups to be tested, and any correction for multiple comparisons, should be fixed before testing.

Comment 4. One run is not the process (footnote 43, paragraphs 35 and 41)

Text concerned. Footnote 43 describes the anonymisation process A as something that can be run on D and on D'.

Problem. When A is generation by a hosted language model, each run gives different output, so one run of A(D) set against one run of A(D') mixes run-to-run variation with the individual's contribution.

Why it matters. Output varies even at settings often described as deterministic, such as a temperature of zero, and the model version behind a service can change or be withdrawn. An assessment of one released output then describes that output, not the process that produced it. The periodic reassessment that paragraph 35 describes as good practice may become impossible if the release cannot be regenerated.

Evidence. On a small sample of respondents, I sent the same request, identical to the character, to the same hosted model version at temperature zero several times. No complete synthetic record came back identical. Several per cent of answers changed from one run to the next, and even whether a run returned a complete record changed. Separately, I could not regenerate a published release. My closest reconstruction differed from the released records clearly more than my own runs differed from one another. I could not tell whether a missing part of the generation recipe or a change in the service caused the gap. Either way, I could not reproduce it from the public material.

Proposed wording (add to footnote 43):

> Where the process A is stochastic, as with generation by a language model, the comparison should use enough independent runs of each version to estimate the variation between runs, the number of runs should be stated, and the difference between versions should be compared with the variation between runs of the same version.

Proposed wording (add to paragraph 41):

> For generated data, the documentation should allow the assessment to be re-run. It should record the model and version identifier, the complete inputs sent to the model, the generation settings, the attributes each record was conditioned on, the outputs actually tested, and the variation measured between repeated runs. An assessment of one generated output should not be assumed to hold for a regenerated output unless it covered several runs.

This extends the documentation that, under Opinion 28/2024 (paragraph 58(e)), supervisory authorities should ideally verify for AI models. The list of conditioning attributes also lets a third party build the matched baseline of Comment 1.

Editorial points

- Footnote 44 refers to “footnote 40” for “prior belief” and to “footnote 41” for A(D'). In version 1.0 these are footnotes 42 and 43.
- The table in paragraph 97 refers to “paras. 98 to 00”. This should read “98 to 100”.
- Paragraph 71(c): “rebuttal presumption” should read “rebuttable presumption”.

Conclusion

The Guidelines already hold the right idea: an inference counts when it could not have been drawn without the person’s data. My four proposals make that idea testable for per-person synthetic data, and each fits in a few sentences of the existing text. I would be glad to provide further technical input, including worked versions of the matched baseline and the planted-trace test.